PAPER NAME

14 Sentimen Analisis Untuk Mengukur Kepercayaan Masyarakat.pdf

AUTHOR

Bagus Setya Rintyarna

WORD COUNT

3388 Words

CHARACTER COUNT

20279 Characters

PAGE COUNT

7 Pages

FILE SIZE

319.8KB

SUBMISSION DATE

Jan 16, 2023 5:00 PM GMT+7

REPORT DATE

Jan 16, 2023 5:00 PM GMT+7

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 7% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 4% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 10 words)
- Manually excluded sources
- Manually excluded text blocks

BIOS : Jurnal Teknologi Informasi dan Rekayasa Komputer
Vol. 3, No. 1, Maret 2022, hlm. 23 - 29
ISSN: 2722-0850
Laman Jurnal: http://bios.sinergis.org

Sentimen Analisis Untuk Mengukur Kepercayaan Masyarakat Terhadap Pengadaan Vaksin COVID-19 Berbasis *Bernoulli Naive Bayes*

Hanifatul Azizah [1], Bagus Setya Rintyarna [2], Triawan Adi Cahyanto [3]
[1]Teknik Informatika, Universitas Muhammadiyah Jember, *hanifatula8@gmail.com*
[2] Teknik Informatika, Universitas Muhammadiyah Jember, *bagus.setya@unmuhjember.ac.id*
[3] Teknik Informatika, Universitas Muhammadiyah Jember, *triawanac@unmuhjember.ac.id*

Keywords:

Bernoulli,
Sentiment,
Analysis,
SMOTE,

ABSTRACT

This study contains an analysis of Indonesian people's sentiments on Twitter towards government policies in handling cases of the COVID-19 pandemic. This study uses the Bernoulli Naive Bayes method in modeling and testing the classification of sentiment data. The performance measurement methods of accuracy, precision and recall are also used to measure the performance of the Bernoulli Naive Bayes method. In the distribution and test scenarios, the K Fold Cross Validation method is used with values of k = 2, 4, 5, 8 and 10. To overcome the data imbalance, in this study the Synthetic Minority Oversampling Technique (SMOTE) technique was used. From the test results with the model without using the Synthetic Minority Oversampling Technique (SMOTE) technique, the results obtained with an accuracy rate of 80.58%, a precision level of 80.33% and a recall rate of 85.57%. while the test results using the Synthetic Minority Oversampling Technique (SMOTE) in modeling, obtained an accuracy rate of 80.20%, a precision level of 78.04% and a recall rate of 86.77%. The test results show that 55% positive sentiment and 45% negative sentiment were obtained using the model without SMOTE, while 53% positive sentiment and 47% negative sentiment were obtained using the model after SMOTE was implemented. The model built without SMOTE implementation has a classification result that is closer to the actual data with a percentage of 58% positive sentiment and 42% negative sentiment.

Kata Kunci

Bernoulli,
Sentimen,
Analisis,
SMOTE,

ABSTRAK

Penelitian ini berisi tentang analisis sentimen masyarakat Indonesia pada Twitter terhadap kebijakan pemerintah dalam menangani kasus pandemi COVID-19. Penelitian ini menggunakan metode *Bernoulli Naive Bayes* dalam melakukan pemodelan dan pengujian klasifikasi terhadap data sentimen. Digunakan juga metode pengukuran performa akurasi, presisi dan *recall* untuk mengukur performa metode *Bernoulli Naive Bayes*. Pada pembagian dan skenario pengujian digunakan teknik *K Fold Cross Validation* dengan nilai k = 2, 4, 5, 8 dan 10. Ketidakseimbangan data dalam penelitian ini diselesaikan dengan menggunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)*. Dari hasil pengujian dengan model tanpa menggunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)* diperoleh hasil dengan tingkat akurasi sebesar 80.58%, tingkat presisi sebesar 80.33% dan tingkat *recall* sebesar 85.57%. sedangkan hasil pengujian dengan menggunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)* pada pemodelan, diperoleh tingkat akurasi 80.20%, tingkat presisi 78.04% dan tingkat *recall* 86.77%. Hasil pengujian menunjukkan diperoleh 55% sentimen positif dan 45% sentimen negatif menggunakan model tanpa *SMOTE* sedangkan dan diperoleh 53% sentimen positif dan 47% sentimen negatif dengan model setelah diimplementasikan *SMOTE*. Model yang dibangun tanpa implementasi *SMOTE* memiliki hasil klasifikasi yang lebih dekat terhadap data aktual dengan persentase 58% sentimen positif dan 42% sentimen negatif.

Korespondensi Penulis:

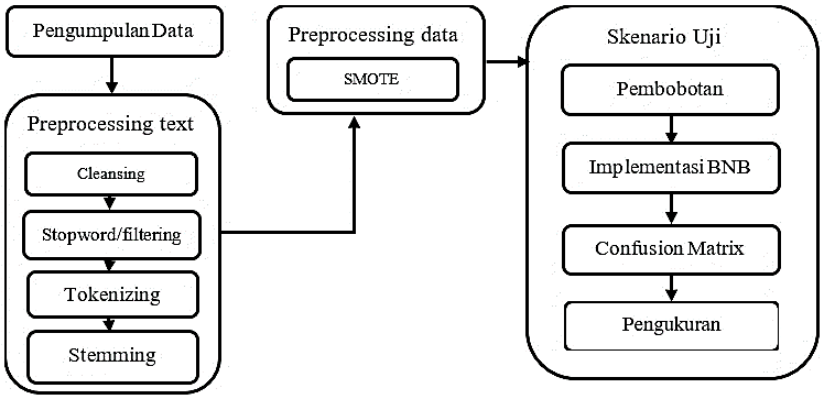
Hanifatul Azizah,
Universitas Muhammadiyah Jember,
Jl. Karimata No 49 Jember
Telepon : +6281338968049
Email: hanifatula8@gmail.com

1. PENDAHULUAN

Pandemik COVID-19 yang terjadi saat ini menyebar dengan sangat cepat, tercatat 91,5 juta kasus di seluruh dunia di antaranya 50,6 juta dinyatakan sembuh dan 1,96 juta dinyatakan meninggal dunia per tanggal 13 Januari 2021 [1]. Meningkatnya penyebaran virus COVID-19 ini membuat pemerintah Indonesia melalui Menteri Kesehatan mengeluarkan surat edaran tentang Protokol Kesehatan pada Situasi Pembatasan Sosial Berskala Besar. Menyikapi pandemi ini banyak warga Indonesia yang masih tidak percaya dan melanggar protokol kesehatan. Di situs resmi Komjen Pol, Wakil Kapolri COVID-19 Polri. Gatot Eddy Pramono mengatakan, pihaknya telah menangkap 5,7 juta orang yang melanggar perjanjian kesehatan di seluruh Indonesia selama seluruh proses pelaksanaan *Health Agreement Implementing Justice* [2].

Pro dan kontra terhadap upaya pemerintah dalam menangani pandemi ini pun banyak ditemui pada *platform-platform* media sosial, salah satunya Twitter. Opini yang dikemukakan masyarakat lewat Twitter terhadap situasi yang terjadi sangat banyak ditemui. Hal ini menuntun penulis untuk melakukan analisis sentimen yang dilontarkan masyarakat pada platform Twitter. Pro dan kontra terhadap virus COVID-19 banyak ditemui di Twitter. Salah satunya pada tulisan dari akun @bayu_r_bay yang menyatakan "*bukan penganut teori konspirasi COVID, cuma paham aja gitu cara bedain bisnis dan bukan, gua percaya COVID itu ada, gua tetap patuhi protokol kesehatan, cuma untuk vaksin gak dulu deh*" [3], mengindikasikan bahwa dia tidak mendukung adanya vaksin COVID-19 dari pemerintah. Tetapi di lain pihak seperti yang dituliskan akun @kataegaa yang menyatakan "*Vaksin dong, ikhtiar. Usaha dulu biar terhindar dari virus, mas. Apa lg dirumah ada orang tua. Kita-kita yg muda dan layak vaksin harus mau, biar Herd Immunity tercapai, jadi bisa melindungi juga yg ggak/belum bisa di vaksin. Dan semoga vaksinnya beneran aman dan efektif*." [4], mengindikasikan bahwa dia mendukung terhadap vaksin COVID-19 dari pemerintah dan banyak lagi sentimen yang dapat kita temui pada platform-platform media sosial lain. Analisis sentimen adalah proses membagi opini menjadi opini positif atau negatif. *Naive Bayes Classifier (NBC)* adalah metode pembelajaran mesin untuk analisis sentimen (Routray et al., 2013). Salah satu algoritma pengembangan dari *Naive Bayes* adalah *Bernoulli Naive Bayes*. Algoritma ini merupakan salah satu teknik klasifikasi yang memiliki keselarasan terhadap penelitian yang diajukan oleh penulis yaitu dengan topik analisis sentimen yang berbasis pada teknik *text mining.*

2. METODE PENELITIAN



Gambar 1. Tahapan Penelitian

2.1 Prosedur Penelitian

Penelitian ini dibagi menjadi beberapa proses di dalamnya. Tiap proses memiliki keterkaitan terhadap peroses setelahnya. Berikut prosedur penelitian yang dibagi menjadi beberapa poin:

1. Pengumpulan Data
   Data dikumpulkan menggunakan teknik crawling pada media sosial Twitter. Pada proses crawling digunakan kata kunci "vaksin", "corona" dan "COVID-19" dalam pengumpulannya. Crawling dilakukan pada tanggal 11 Januari 2021 sampai 3 Februari 2021.
2. *Preprocessing*
   Preprocessing yang dilakukan pada penelitian ini antara lain [5]:
   a. *Cleansing*, atau pembersihan data dari angka, karakter asing, link dan lainnya selain huruf.
   b. *Stopword removal* atau penghapusan kata yang tidak memiliki pengaruh pada sentimen.
   c. *Tokenizing* atau pemecahan kalimat menjadi per kata.
   d. *Stemming* atau pengembalian kata ke kata dasar [6].

3. Pembobotan

Teknik pembobotan yang digunakan adalah *TF-IDF*. *Term frequency* adalah metode penghitungan bobot setiap kata dalam sebuah teks. Dalam metode ini, diasumsikan bahwa nilai kepentingan setiap istilah sebanding dengan berapa kali istilah tersebut muncul dalam teks [7]. *Inverse document frequency (IDF)* yang mengurangi dominasi istilah-istilah yang kerap ditemukan di berbagai data teks. Kondisi ini ditujukan untuk menemukan kata penting yang memiliki frekuensi kemunculan kecil [8].

$$idf_j = \log\left(\frac{D}{df_j}\right) \quad (1)$$

$$TFIDF(d,t) = TF(d,t).IDF(t) \quad (2)$$

Dimana, istilah frekuensi dari suku *t* dalam teks *d*. *D* adalah jumlah semua data teks yang diperoleh dan df_j adalah jumlah data teks yang memiliki kata *t j*.

4. *Imbalanced Data*

Kondisi data yang tidak seimbang dapat mempengaruhi hasil klasifikasi nantinya. Untuk mengatasi ini digunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)*. Teknik ini bekerja dengan cara membangun data sintetis dari data minoritas agar seimbang terhadap data mayoritas. *Synthetic Minority Oversampling Technique (SMOTE)* merupakan metode pengembangan dari *oversampling*. *SMOTE* dikemukakan oleh oleh Nithes V. Chawla. Metode ini bekerja dengan membuat beberapa salinan data. Jenis penyalinan ini disebut data sintetis. Penerapan *SMOTE* untuk meminimalkan ketidakseimbangan kelas sehingga memiliki model yang baik [9]. Teknik ini bekerja dengan cara membuat data sintetis atau data buatan berdasarkan pengukuran kedekatan data numerik dengan jarak *Euclidean*, sedangkan data klasifikasinya lebih sederhana yaitu nilai modus. Berikut persamaan yang digunakan:

$$X_{syn} = X_i + (X_{knn} - X_i)\ x\ \delta \quad (3)$$

Dimana, X_{syn} adalah data sintesis yang akan diciptakan, X_{knn} merupakan jarak terdekat dari data yang dibuat sintesisnya, X_i adalah data dengan atribut ke-*i* dan δ nilai random antara 0 dan 1.

5. Pemodelan

Pemodelan yang digunakan menggunkan metode *Bernoulli Naive Bayes* dengan menggunakan skenario uji yang dibantu oleh *K Fold Cross Validation* untuk menemukan hasil pengukuran terbaik. Data yang terdiri dari sentimen positif dan negatif dipartisi menjadi dua bagian yaitu data latih dan data uji validasi. Ke dua partisi ini memiliki persentase 80% untuk data latih dan 20% untuk data uji validasi. 80% data latih akan dilakukan pemodelan serta pengukuran performa dengan dibantu menggunkaan skenario K Fold Cross Validation guna memperoleh nilai performa tertinggi dari data latih tersebut. Sedangkan 20% data uji validasi digunakan untuk mengukur model yang telah dibangun pada data latih tersebut. Pemodelan ini juga akan memanfaatkan *Synthetic Minority Oversampling Technique (SMOTE)* untuk mengatasi ketidakseimbangan data. Terdapat dua model, yaitu model tanpa implementasi *SMOTE* pada data latih dan implementasi menggunakan *SMOTE* pada data latih.

Teori klasifikasi Bayesian adalah metode statistik dasar dari *data mining*. Metode ini didasarkan pada penggunaan probabilitas untuk menghitung *trade-off* antara berbagai keputusan klasifikasi [10]. Metode *Naive Bayes* memiliki beberapa algoritma khusus dalam distribusinya, di antaranya Gaussian *Naive Bayes*, *Bernoulli Naive Bayes* dan Polynomial *Naive Bayes*. Algoritma *Bernoulli Naive Bayes* mengklasifikasikan distribusi data menurut distribusi multivariate *Bernoulli*. Aturan keputusan untuk algoritma *Bernoulli Naive Bayes* adalah sebagai berikut:

$$P(A_i|B) = P(i|B)A_i + (1 - P(i|B))(1 - A_i) \quad (4)$$

Skenario uji pada penelitian ini akan menggunakan *K Fold Cross Validation* dengan k subset = 2, 4, 5, 8 dan 10. *K Fold Cross Validation* adalah cara yang dapat digunakan untuk menemukan hasil pengukuran terbaik dengan cara melakukan pengujian silang pada suatu data. Cara ini bekerja dengan cara membagi kumpulan data menjadi k subset atau bagian. Jika salah satu subset bertindak sebagai bahan uji, maka subset k-1 lainnya bertindak sebagai bahan latih. Proses ini dijalankan *k* kali, sehingga setiap subset akan menjadi data uji model [11].

2.2 Prosedur Pengujian

Pengujian pada penelitian ini akan menggunakan metode yang sama yaitu *Bernoulli Naive Bayes* dan hasil klasifikasi dihitung menggunkaan *Confusion Matrix* [12].

		Klasifikasi	
Aktual		TP	FN
		FP	TN

Gambar 2. *Confusion Matrix* 2 Classes

Dimana, FN merupakan *false negative*, TP merupakan *true positive*, TN merupakan *true negative* sedangkan FP merupakan *false positive* dan. Keempat kriteria ini akan digunakan dalam mengukur performa metode yang digunakan. Dalam penelitian ini akan diukur tingkat akurasi dan presisi. Berikut persamaan dari keduanya:

$$Presisi = \frac{TP}{TP+FP} \quad (5)$$

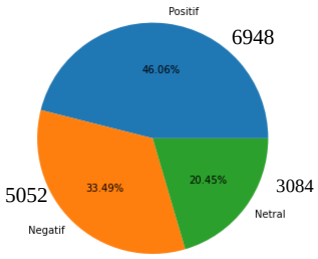
$$Akurasi = \frac{TP+FP}{TP+FP+TN+FN} \quad (6)$$

3. HASIL DAN ANALISIS

Berikut hasil analisis pada penelitian yang dilakukan. Hasil dan analisis terbagi menjadi empat sub bab sebagai berikut:

3.1 Gambaran data

Data yang diperoleh dari hasil *crawling* berjumlah 15.084 *tweet*. Data selanjutnya mendapat validasi oleh tenaga ahli untuk menentukan nilai aktual. Berikut sebaran data pada data keseluruhan. Penelitian ini menggunakan data dengan sentimen positif dan negatif yang akan diteliti atau tidak menggunkaan data dengan sentimen netral. Dari hasil validasi oleh pihak ahli, dalam penelitian ini dibantu oleh dosen Bahasa Indonesia untuk menentukan nilai sentimen sebagai bahan validasi. Sehingga total data yang akan dilakukan penelitian berjumlah 12.000 yang terdiri dari sentimen negatif dan sentimen positif. Potongan data hasil validasi tenaga ahli dijelaskan pada tabel 1 di bawah ini.



Gambar 2. Hasil pengumpulan data keseluruhan

Penelitian ini akan menggunakan data dengan sentimen positif dan negatif, maka data netral tidak akan digunakan. Sehingga total data menjadi 12.000 *tweet* dengan kategori 5.052 *tweet* negatif dan 6.948 *tweet* positif.

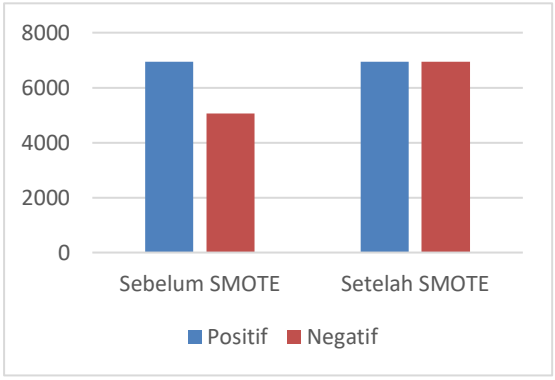
Tabel 1. Potongan data hasil validasi oleh tenaga ahli

No.	screen_name	text	sentimen
0	mwahyuarif_	Alhamdulillah, Pak Presiden resmi sudah divaks...	p
1	lovelyminghaoo	@wonwxzy loh emngnya kenapa? corona itu penyak...	p
2	DamailahN	@PolisiSukoharjo Semoga vaksinasi menjadi pera...	p
3	BlackBo09654232	@PolisiSukoharjo Korban jiwa telah banyak akib...	n
4	AchmadSukarno5	@jokowi Semoga lancar,dan sesuai apa yg kita h...	p
5	NaufalIvander1	@avantwijaya051 @jokowi Corona muncul -> Mi...	o
...	...	...	...
15079	hariethzhasnin	@luqmannaj Vaksin tak percaya, Luqman comel pe...	p
15080	Schupreme	vaksin cepetin napa, gua pengen ngekos	p
15081	aesdeefgeha	@padangmenfess Banyak nder, bahas vaksin COVID...	p
15082	INUM4KII	Abang jadinya bolak balik wkwkkw hari ini pula...	p
15083	balistories	Seratusan SDMK di Tabanan Tak Penuhi Syarat Pe...	p

Tabel 1 di atas merupakan hasil validasi oleh tenaga ahli pada data keseluruhan. Pada tabel tersebut nilai sentimen positif dilambangkan menggunakan huruf “p”, nilai sentimen negatif dilambangkan menggunakan huruf “n” dan nilai sentimen netral dilambangkan menggunkaan huruf “o”.

3.2 Hasil Implementasi *SMOTE*

Hasil data yang terkumpul dan digunakan dalam penelitian ini terdiri dari 5.052 *tweet* yang bernilai negatif dan 6.948 *tweet* yang bernilai positif. Kondisi ini yang dinilai kurang seimbang. *SMOTE* digunakan untuk menyeimbangkan kondisi ini. Berikut digambarkan hasil implementasi *SMOTE* pada data latih dan kondisi data latih sebelum diimplementasi *SMOTE*.



Gambar 3. Kondisi data latih sebelum diimplementasi *SMOTE* (kiri) dan kondisi data latih setelah diimplementasi *SMOTE* (kanan)

Kondisi pada Gambar 3 grafik sebelah kiri merupakan data latih asli tanpa implementasi teknik *SMOTE*. Kondisi dengan sentimen positif yang memiliki jumlah lebih banyak dari data sentimen negatif serta grafik sebelah kanan merupakan data latih setelah dilakukan implementasi *SMOTE*, dimana kondisi data dengan sentimen negatif memiliki jumlah yang sama dengan jumlah sentimen positif.

3.3 Hasil Pemodelan Pada *K Fold Cross Validation*

Pemodelan dilakukan dua versi, pertama dilakukan pemodelan terhadap data latih tanpa menggunakan teknik *SMOTE* dan kedua dilakukan pemodelan terhadap data latih setelah diimplementasikan teknik *SMOTE*. Berikut hasil pemodelan yang dijelaskan pada tiap poin:

1. Pemodelan pada data latih tanpa implementasi *SMOTE*
   Pemodelan pertama dalam penelitian ini dilakukan tanpa implementasi teknik *SMOTE* pada data. Berikut hasil dari pengukuran hasil model yang diperoleh yang dipaparkan pada tabel di bawah ini.

Tabel 1. Hasil pemodelan menggunakan skenario *K Fold Cross Validation* pada data latih tanpa implementasi *SMOTE*

	Rata-Rata		
Nilai K	**Akurasi**	**Presisi**	***Recall***
2	80.16	79.66	80
4	80.11	79.65	80.1
5	80.08	79.64	80.1
8	80.12	79.67	80.13
10	80.22	79.77	80.24

Dari hasil di atas, rata-rata pada tingkat akurasi yang diperoleh cenderung sama. Hal serupa juga diperoleh pada pengukuran tingkat presisi dan *recall*.

2. Pemodelan pada data latih dengan implementasi teknik *SMOTE*
   Pemodelan kedua dalam penelitian ini dilakukan tanpa implementasi teknik *SMOTE* pada data. Berikut hasil dari pengukuran hasil model yang diperoleh yang dipaparkan pada tabel di bawah ini.

Tabel 2. Hasil pemodelan menggunakan skenario *K Fold Cross Validation* pada data latih dengan implementasi *SMOTE*

	Rata-Rata		
Nilai K	**Akurasi**	**Presisi**	***Recall***
2	81.48	81.71	81.48
4	81.64	81.87	81.65
5	81.66	81.92	81.66
8	81.81	82.09	81.81
10	81.83	82.12	81.83

Berdasarkan tabel di atas, hasil pemodelan setelah *SMOTE* memiliki tingkat akurasi, presisi dan *recall* lebih tinggi. Kondisi tiap k juga memiliki konsistensi yang hampir sama atau tidak jauh beda.

Dari tabel 4.18 dapat dijelaskan bahwa hasil pengukuran tingkat akurasi pada percobaan menggunakan teknik *SMOTE* memiliki nilai akurasi lebih baik. Peningkatan nilai akurasi pada percobaan menggunakan *SMOTE* rata-rata sebesar 1.55% atau dalam hal ini hasil percobaan dengan implementasi *SMOTE* lebih baik dari percobaan tanpa *SMOTE*. Rata-rata tingkat akurasi pada percobaan tanpa *SMOTE* adalah 80.14% dan rata-rata tingkat akurasi percobaan menggunakan *SMOTE* adalah 81.68%.

Pada pengukuran tingkat presisi dapat dijelaskan bahwa hasil pengukuran tingkat presisi pada percobaan menggunakan teknik *SMOTE* memiliki nilai presisi lebih baik. Peningkatan nilai presisi pada percobaan menggunakan *SMOTE* rata-rata sebesar 2.26% atau dalam hal ini hasil percobaan dengan implementasi *SMOTE* lebih baik dari percobaan tanpa *SMOTE*. Rata-rata tingkat presisi pada percobaan tanpa *SMOTE* adalah 79.68% dan rata-rata tingkat presisi percobaan menggunakan *SMOTE* adalah 81.94%.

Pada pengukuran tingkat *recall* dapat dijelaskan bahwa hasil pengukuran tingkat *recall* pada percobaan menggunakan teknik *SMOTE* memiliki nilai *recall* lebih baik. Peningkatan nilai *recall* pada percobaan menggunakan *SMOTE* rata-rata sebesar 1.57% atau dalam hal ini hasil percobaan dengan implementasi *SMOTE* lebih baik dari percobaan tanpa *SMOTE*. Rata-rata tingkat *recall* pada percobaan tanpa *SMOTE* adalah 80.11% dan rata-rata tingkat *recall* percobaan menggunakan *SMOTE* adalah 81.69%.

3.4 Hasil Pengujian

Pengujian dilakukan pada data *tweet* uji validasi pada porsi 20% dari total data. Berikut hasil pengukuran pada pengujian data sebelum dan setelah diimplementasikan teknik *SMOTE*.

Tabel 3. Hasil uji validasi pada data uji

Skenario	Akurasi	Presisi	*Recall*
tanpa *SMOTE*	80.58	80.33	85.57
dengan *SMOTE*	80.2	78.04	86.77

Berdasarkan tabel di atas, data sebelum *SMOTE* memiliki tingkat akurasi dan presisi lebih tinggi. Ini berbanding terbalik dengan hasil pemodelan yang diperoleh dimana data dengan teknik *SMOTE* memperoleh model yang lebih baik. Pada tabel di atas juga dijelaskan bahwa terjadi penurunan performa pada data dengan implementasi *SMOTE*.

4. KESIMPULAN

Dari hasil penelitian yang dilakukan, diperoleh beberapa poin kesimpulan yang dapat dijelaskan sebagai berikut ini:

a. Hasil implementasi *Bernoulli Naive Bayes* terhadap data uji validasi tanpa menggunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)* yaitu diperoleh persentase tingkat sentimen positif sebesar 55% atau 1317 data dari 2400 data. Sedangkan pada uji coba menggunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)* diperoleh nilai persentase sentimen positif sebesar 53% atau 1261 data dari 2400 data.
b. Hasil implementasi *Bernoulli Naive Bayes* terhadap data uji validasi tanpa menggunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)* yaitu diperoleh persentase tingkat sentimen negatif sebesar 45% atau 1083 data dari 2400 data. Sedangkan pada uji coba menggunakan teknik *Synthetic Minority Oversampling Technique (SMOTE)* diperoleh nilai persentase sentimen negatif sebesar 47% atau 1139 data dari 2400 data.
c. Nilai akurasi, presisi dan *recall* yang dihasilkan metode *Bernoulli Naive Bayes* dalam mengklasifikasi data sentimen dari Twitter terhadap data uji validasi model pada percobaan tanpa *Synthetic Minority Oversampling Technique (SMOTE)* yaitu memiliki tingkat akurasi 80.58%, presisi 80.33% dan 85.57% sedangkan pada percobaan menggunakan tanpa *Synthetic Minority Oversampling Technique (SMOTE)* yaitu memiliki tingkat akurasi 80.20%, presisi 78.04% dan *recall* 86.77%
d. Percobaan implementasi *Bernoulli Naive Bayes* menggunakan *Synthetic Minority Oversampling Technique (SMOTE)* memiliki nilai akurasi, presisi dan *recall* lebih tinggi pada skenario uji *Cross Fold Validation* tetapi memiliki nilai akurasi dan presisi lebih kecil pada uji data validasi.

UCAPAN TERIMA KASIH

Ucapan terima kasih diperuntukkan bagi Universitas Muhammadiyah Jember dan semua pihak yang telah membantu dalam penelitian ini.

REFERENSI

[1] JOHNS HOPKINS WHITING SCHOOL OF ENGINEERING, "JHU CSSE – Center For Systems Science and Engineering at JHU," *www.ystems.jhu.edu*, 2020. https://systems.jhu.edu/ (accessed May 20, 2020).

[2] Www.covid19.go.id, "Terapkan Protokol Kesehatan, Polisi Jaring 5,7 Juta Pelanggar," *www.covid19.go.id*, 2020. https://covid19.go.id/p/berita/terapkan-protokol-kesehatan-polisi-jaring-57-juta-pelanggar

[3] M. Pluto, "Tweet: bukan penganut teori konspirasi COVID, cuma paham...," *www.twitter.com*, 2021. https://twitter.com/bayu_r_bay/status/1349240281607012354

[4] Ega, "Tweet : Vaksin dong, ikhtiar. Usaha dulu biar terhindar dari virus...," *www.twitter.com*, 2021. https://twitter.com/kataegaa/status/1349250373022674945

[5] E. Nugroho, *Perancangan Sistem Deteksi Plagiarisme Dokumen Teks Dengan Menggunakan Algoritma Rabin-Karp*, vol. 34, no. 2. Universitas Brawiaya, 2011. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01402390.2011.569130%5Cnhttp://proxy.library.upenn.edu:2195/doi/abs/10.1080/01402390.2011.569130

[6] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.

[7] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper," *FLAIRS Conf.*, pp. 235–239, 1995.

[8] I. H. Witten, Z. Bray, M. Mahoui, and B. Teahan, "Text mining: a new frontier for lossless compression," *Data Compression Conf. Proc.*, pp. 198–207, 1999, doi: 10.1109/dcc.1999.755669.

[9] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, 2018.

[10] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*, no. May. Informatika Bandung, 2017.

[11] T. Rosandy, "PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4.5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus : KSPPS / BMT AL-FADHILA," *J. Teknol. Inf. Magister Darmajaya*, vol. 2, no. 01, pp. 52–62, 2016.

[12] M. Haltuf, "Support Vector Machines for Credit Scoring," no. August 2014, 2014.

10% Overall Similarity

Top sources found in the following databases:

- 7% Internet database
- Crossref database
- 4% Submitted Works database
- 2% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	repository.unmuhjember.ac.id Internet	2%
2	libguides.niu.edu Internet	1%
3	media.neliti.com Internet	1%
4	I Ariyati, S Rosyida, K Ramanda, V Riyanto, S Faizah, Ridwansyah. "Opti... Crossref	<1%
5	id.123dok.com Internet	<1%
6	Universitas Gunadarma on 2021-01-07 Submitted works	<1%
7	University of Nottingham on 2020-05-14 Submitted works	<1%
8	etheses.uin-malang.ac.id Internet	<1%

9	ejournal.stitpn.ac.id Internet	<1%
10	"Technology for Smart Futures", Springer Science and Business Media ... Crossref	<1%
11	UIN Walisongo on 2022-11-07 Submitted works	<1%
12	journal.ummat.ac.id Internet	<1%
13	Sriwijaya University on 2022-01-20 Submitted works	<1%
14	Universitas Brawijaya on 2018-01-24 Submitted works	<1%
15	journal.thamrin.ac.id Internet	<1%
16	repository.uin-suska.ac.id Internet	<1%